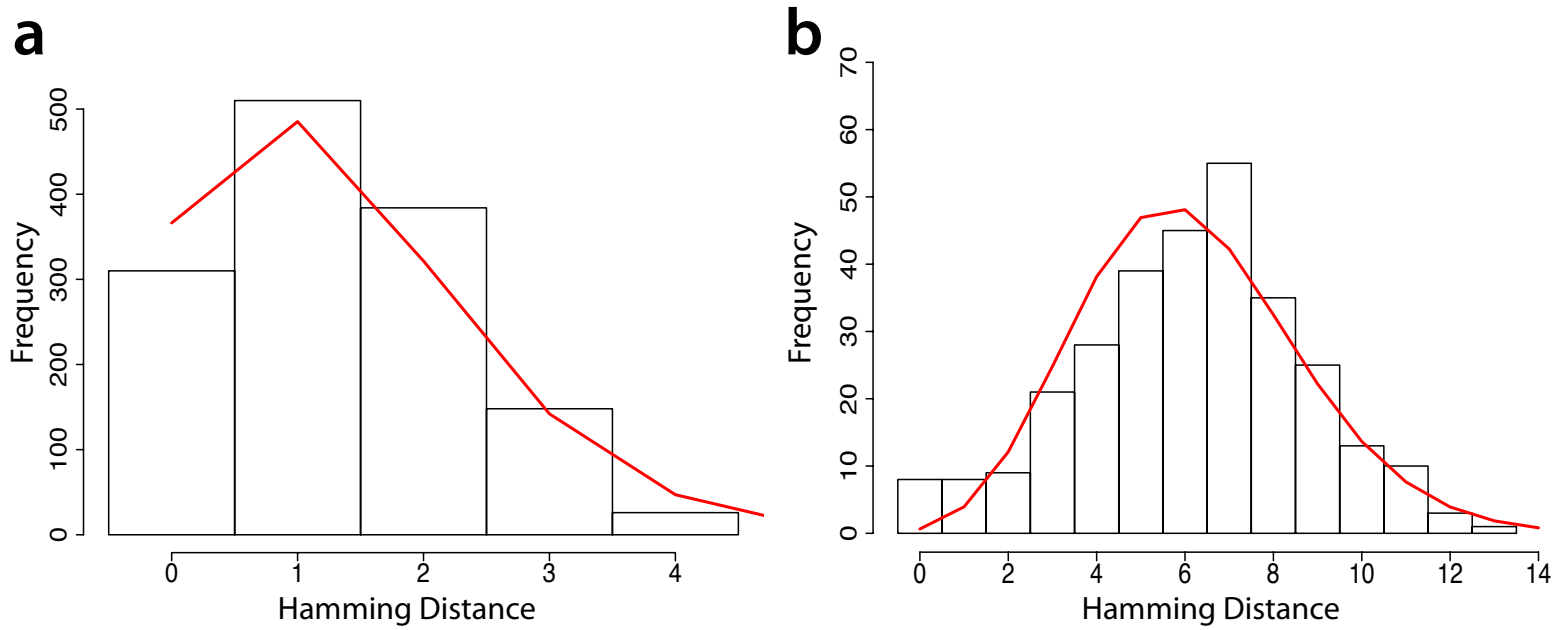
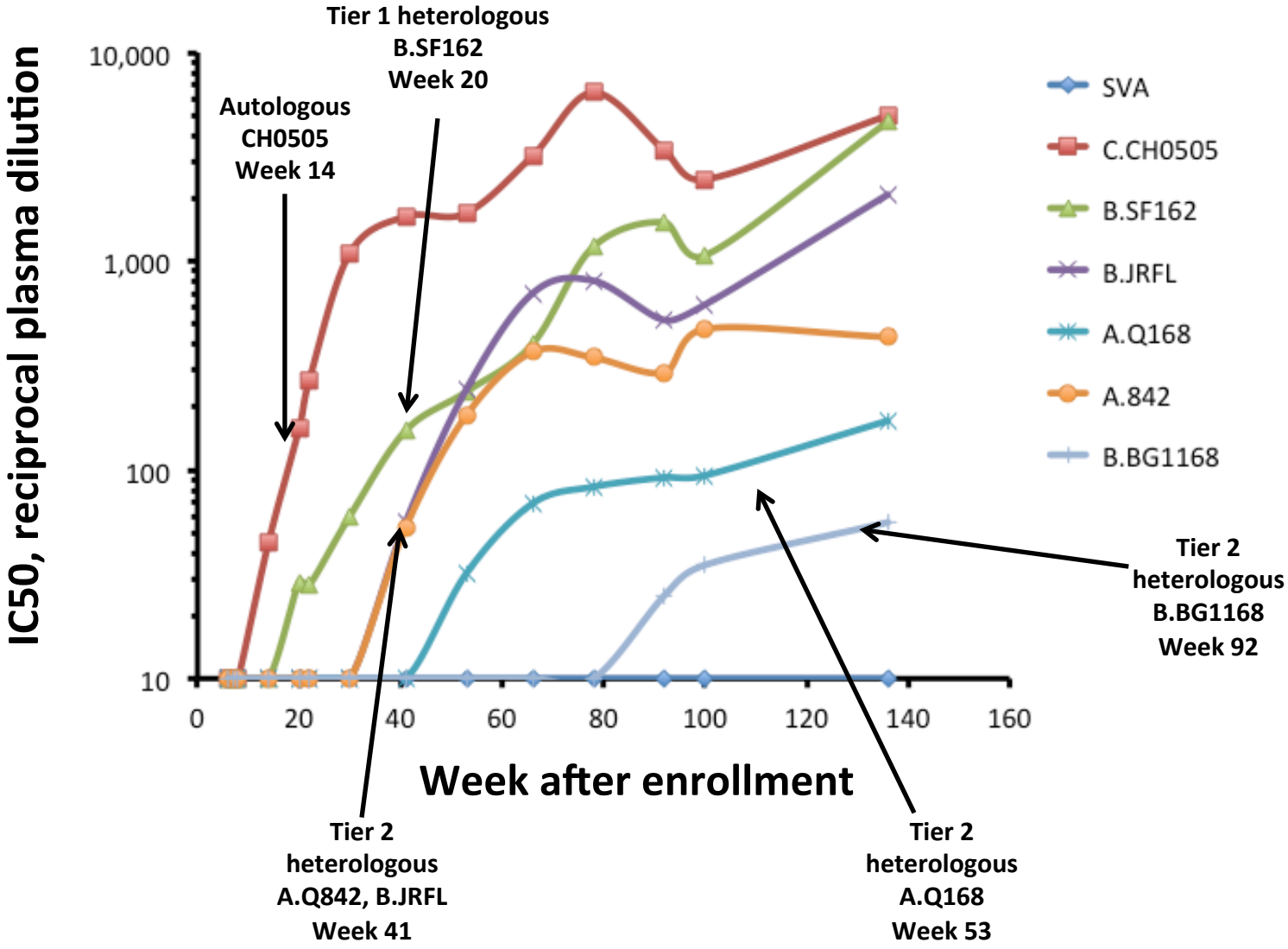
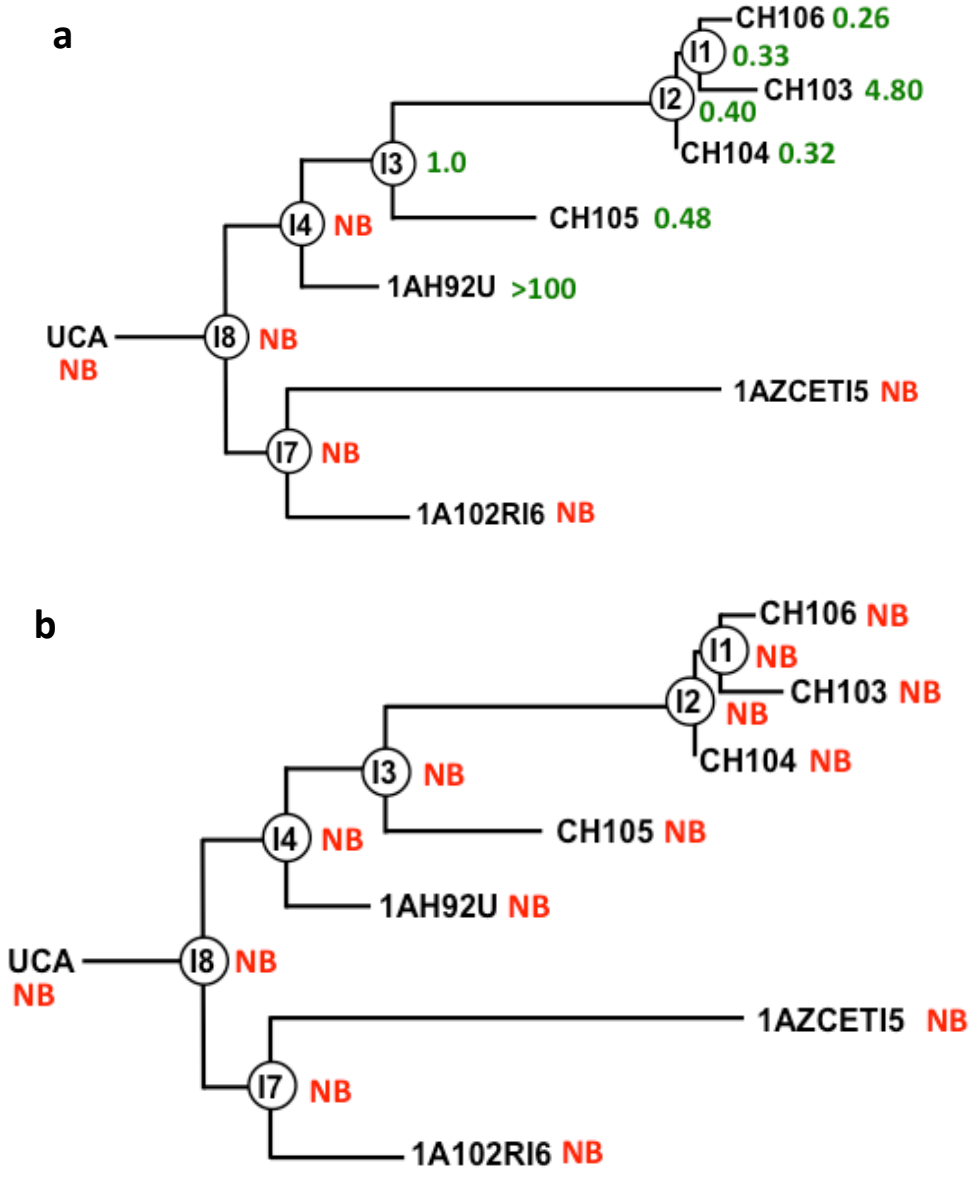


Supplementary Figure 1. Hamming distance frequency distributions of sequences at (a) week 4 and (b) week 14. A model of the best fit Poisson distribution is shown as a red line. Analysis of the sequence diversity in the first available sample (a) from subject CH505 using the Poisson Fitter tool (ref below) indicates that the sequences were a consistent with a star phylogeny and that the mutations were accumulating according to a Poisson distribution (goodness of fit p = 0.11). This is consistent with a single founder virus establishing the infection, with random accumulation of mutations prior to selection. The lambda parameter was 1.325, and assuming the mutation rate of 2.16 10^{-05}, the estimated time from the most recent common ancestor was 22 days (95% CI, 18-27). Given that the outer bound of this confidence interval is 27 days, it is highly like this sample was taken within 4 weeks of infection, thus we are calling this sampling time "week 4" as a conservative estimate. This timing estimate is further supported by Feibig staging at time of enrollment. By week 14 (b), the tree was no longer consistent with a star phylogeny or a Poisson distribution ($p << 10^{-10}$), indicating selection was well underway. Of note, although the mutation data at week 4 (a) is statistically consistent with a Poisson distribution, the observed number of pairwise sequence identities was somewhat reduced relative to expectation, and the observed number of Hamming distances of 1 and 2 are slightly more than expected. This is of interest as this shift is the a result of a single mutation in loop D, in a CH103 contact residue (N279K) -- so although the deviation from the Poisson was not significant, given its location it is possible that the site is a very early indicator of selection.

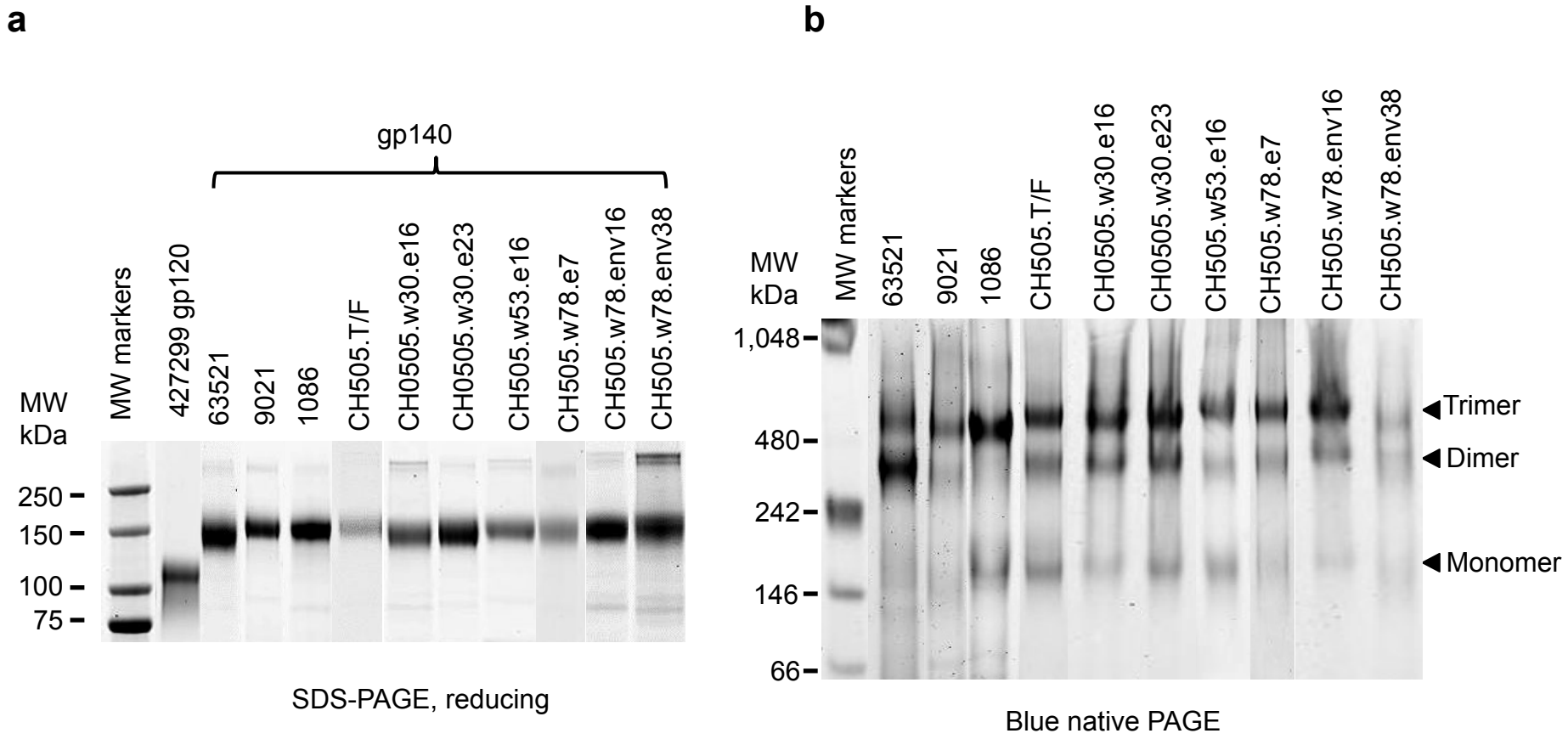
Reference: Giorgi EE, Funkhouser B, Athreya G, Perelson AS, Korber BT, Bhattacharya T. Estimating time since infection in early homogeneous HIV-1 samples using a Poisson model. BMC Bioinformatics 2010 Oct 25;11:532. PMID: 20973976
http://www.hiv.lanl.gov/content/sequence/POISSON_FITTER/poisson_fitter.html



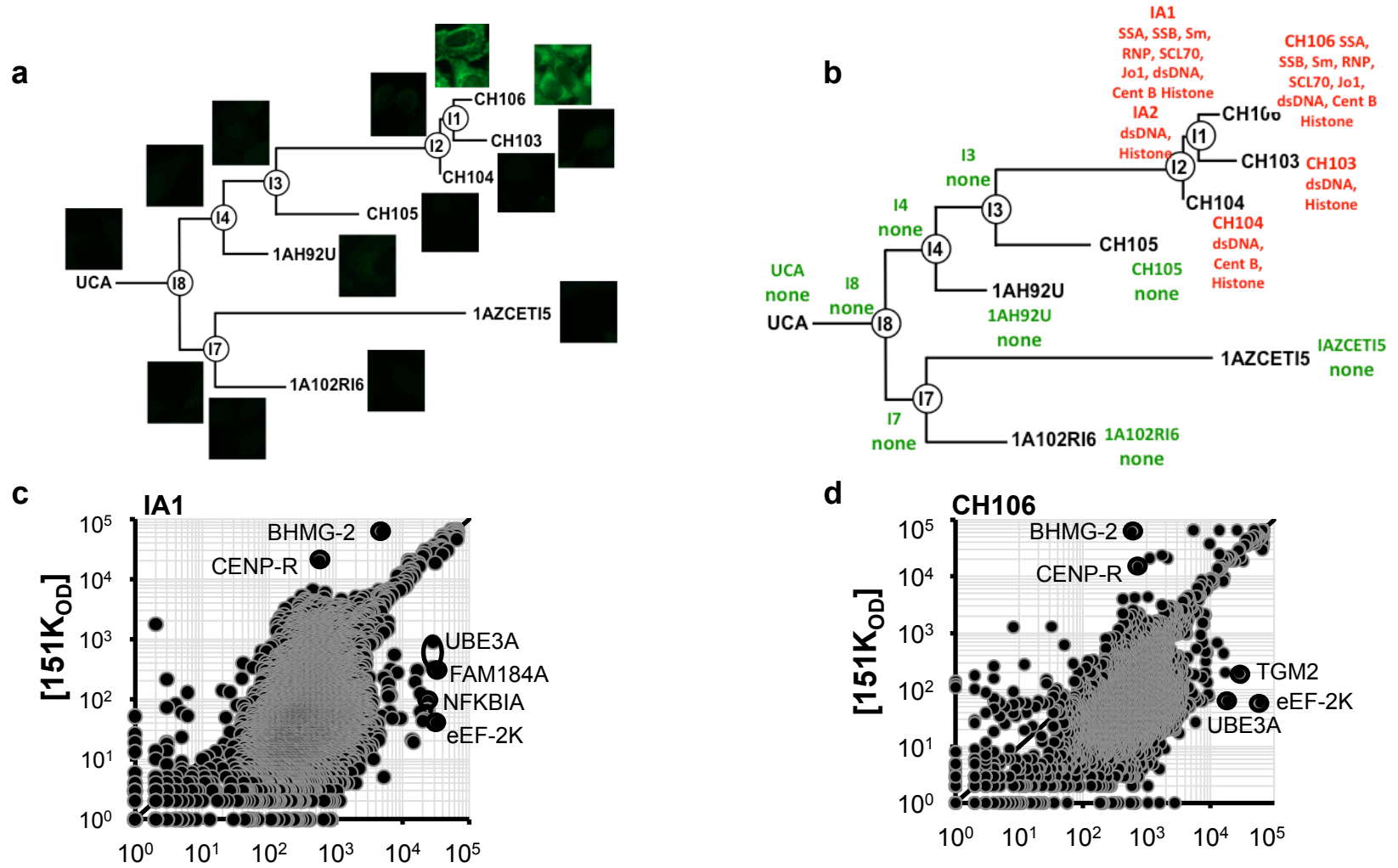
Supplementary Figure 2. Plasma neutralization of CH505 patient over time to autologous transmitted/founder (T/F) and heterologous HIV-1 viruses. Plasma samples were longitudinally collected from HIV-1 patient CH505 starting from time of infection(in x axis) and tested for neutralization activity against the autologous transmitted/founder (T/F) virus and heterologous HIV-1 Env pseudoviruses including subtype B (B) SF162, JRFL and BG1168) and subtype A (Q168 and 842) in TZM-bl cell-based neutralization assays. Results were expressed as IC50 (reciprocal plasma dilution) (in y axis).



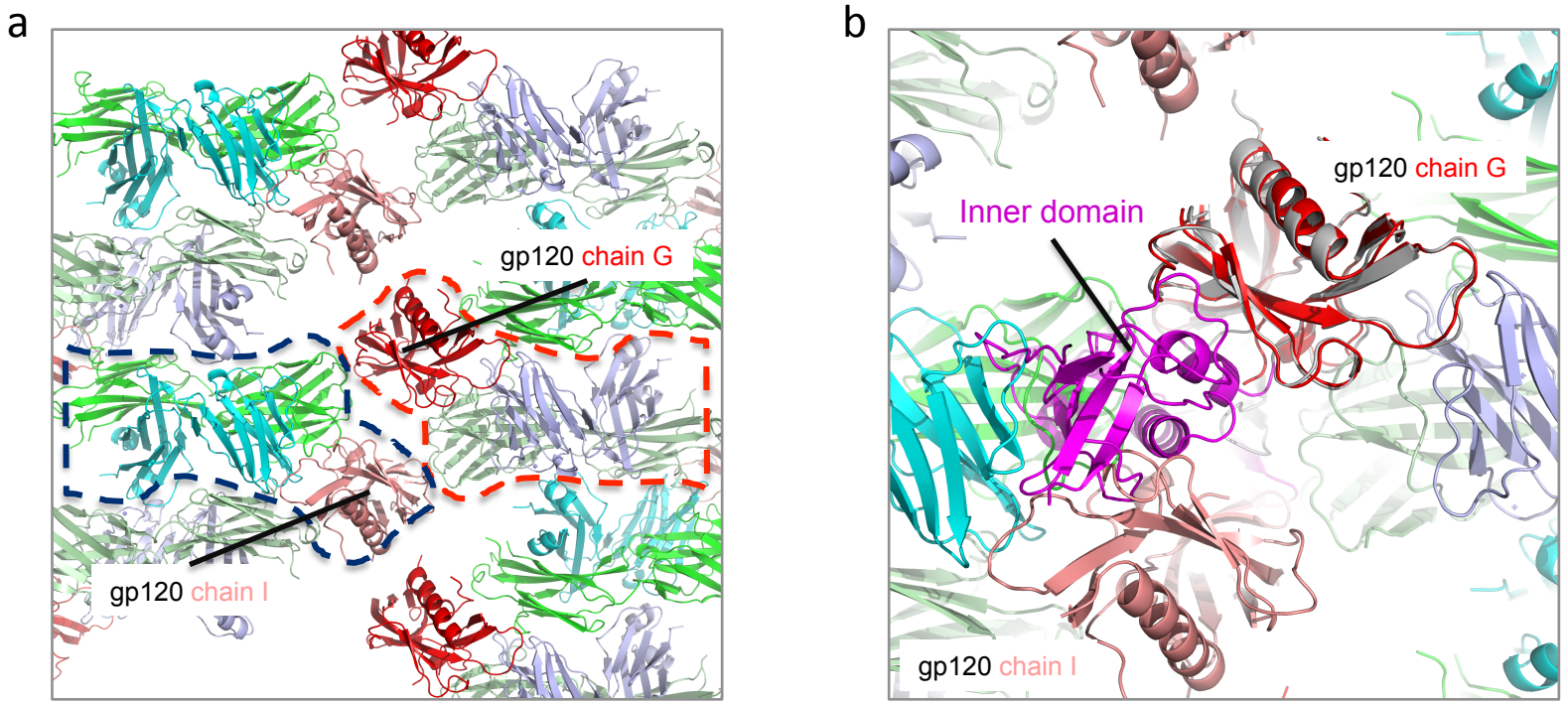
Supplementary Figure 3. Reactivity of antibodies in CH103 clonal lineage to HIV-1 Env resurfaced core3 (RSC3) and RSC3 mutant. Antibodies in CH103 clonal lineage were tested in dose range from 100ug to 0.0005ug/ml for binding to (**a**) HIV-1 Env RSC3 and (**b**) RSC3 mutant with P363N and D371I mutations in ELISA. Results are expressed as EC50 (ug/ml) and are indicated next to the individual antibodies. NB = no detectable binding.



Supplementary Figure 4. SDS-PAGE analysis of recombinant HIV-1 Env gp140 and gp120 proteins. HIV-1 Env gp140 and gp120 proteins were analyzed on SDS-PAG under reducing condition (**a**) and gp140 proteins were analyzed on blue negative PAGE (**b**). Individual HIV-1 Env proteins are identified on the top of gels. **a,** The HIV-1 gp120 and gp140 used in the study had no degradation under reducing condition in SDS-PAGE. **b,** Most heterologous HIV-1 Env gp140 Envs and all autologous CH505 gp140 Envs migrated predominantly as trimers and also contain dimer and monomer forms.

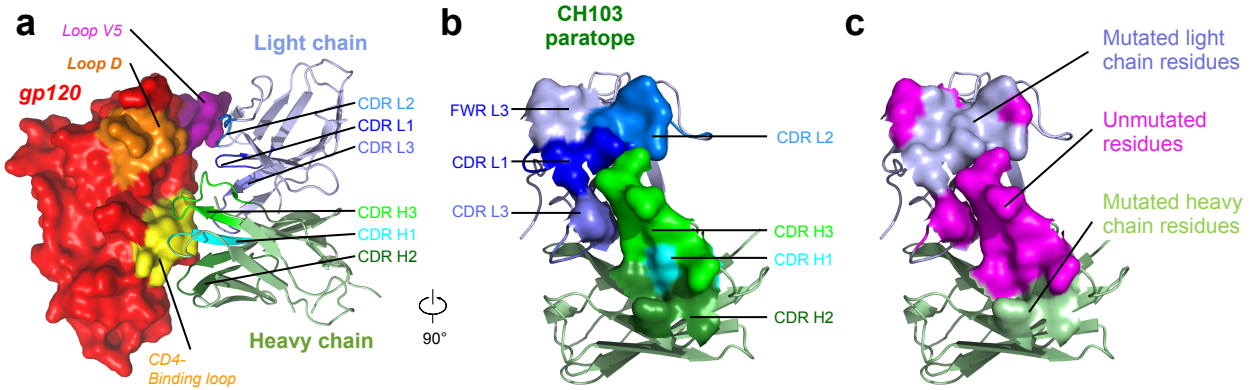


Supplementary Figure 5. Polyreactivity analysis of antibodies in CH103 clonal lineage by HEp-2 staining, ANA assays and protein array microchip analysis. Reactivity of antibodies in CH103 clonal lineage was assayed by indirect immunofluorescence HEp-2 staining (**a**) and by ANA assays (**b**). Pictures at magnification x 200 of immunofluorescence staining for individual antibodies are presented next to the antibody ID. Results of the reactivity of individual antibodies with panel of autoantigens assayed by ANA are indicated (**b**). The intermediate antibody (I1) and CH106 were identified as reactive with HEp-2 cells and then selected for further testing for reactivity with human host cellular antigens (**c** and **d**) using Invitrogen ProtoArrays^{TM} . We found that I1 (**c**) and CH106 ((**d**) exhibit specific autoreactivity and robust polyreactivity. Bound antibody was determined by immunofluorescence and relative fluorescence intensities for 9,400 recombinant human proteins in the 151K array (y-axis) that were plotted against (x-axis) the homologous intensities in IA1 (**c**) and CH106 (**d**) arrays. All proteins were printed in duplicate on each array and each data point represents one fluorescence measurement. The diagonals in each graph represent equal fluorescence intensities (equivalent binding) by the I1, CH106 and 151K control Ab. Self-antigens bound by the I1 and CH106 are identified by high fluorescence intensity versus 151K and are indicated by circles. Polyreactivity was indicated by significant and general skewing from the diagonal. Autoantigens identified: BHMT2 (betaine-homocysteine methyltransferase 2); CENP-R (centromere protein R) [151K]; eEF-2K (eukaryotic elongation factor-2 kinase); UBE3A (ubiquitin-protein ligase E3A) [IA1 and CH106]; TGM2 (transglutaminase 2) [CH106]; NFKBIA (nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha); FAM184A (family with sequence similarity 184, member A) [I1].



Supplementary Figure 6. Crystal packing of the CH103-gp120 complex in $P2_1$ space group. a, A view of the crystal lattice. The two complexes in each asymmetric unit are marked with red and blue dashed lines and are shown in cartoon diagrams with gp120 in red and salmon, CH103 heavy chain in green and palegreen, and CH103 light chain in light blue and cyan. **b**, A close-up view of the lattice between two neighboring complexes. When extended core gp120 of clade C ZM176.66 from the VRC01 complex is superposed with its ordered corresponding portions in the CH103 complex, the inner domain shown in magenta clashes with the neighboring complex, indicating inner domain of gp120 is not present in the CH103-gp120 crystal due to proteolytic degradation during crystal growth.

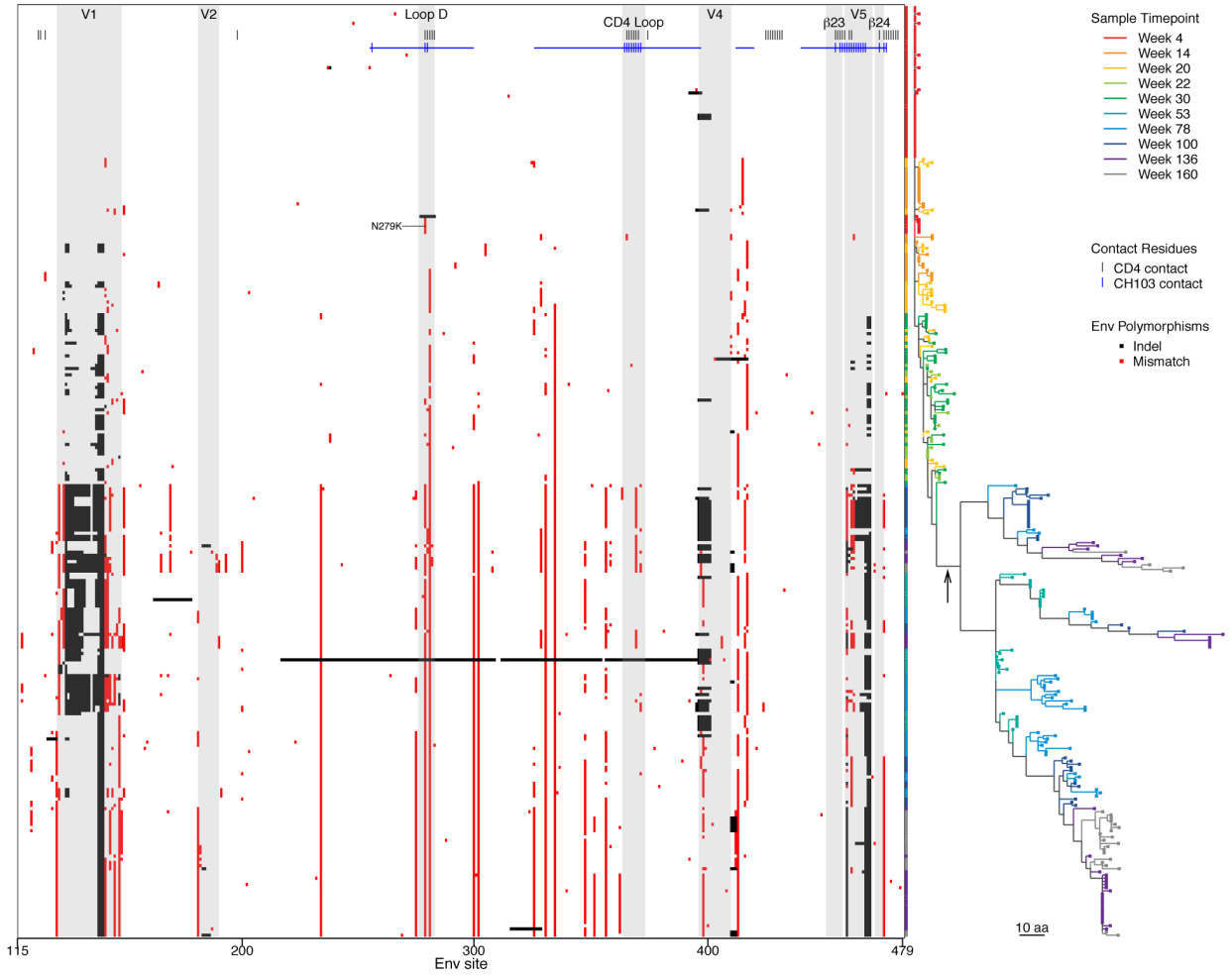
Figure 7



d

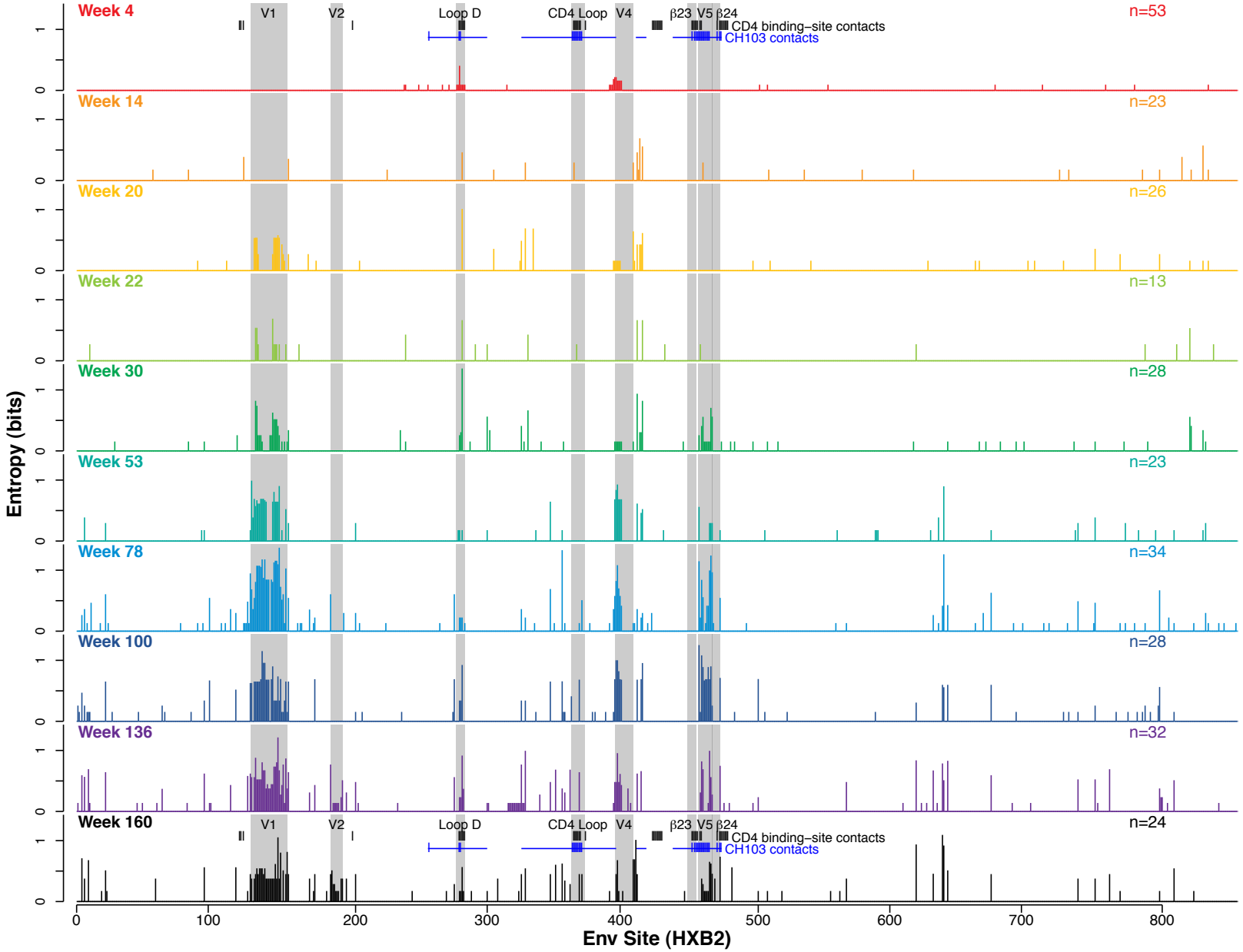
```
Heavy chain
                1                   22                  36                  50                  66                 82abc                92                 103
                |                   |                   |                   |                   |                   |                   |                   |
                ----------FWR1------------_CDR1-----FWR2-----_____CDR2_____----------FWR3-----------_____CDR3_____----FWR4---
                                          ☼                 ☼ ☼ ☼ ● ☼                                   ☼○●●●☼
UCA             QVQLQESGPGLVKPSETLSLTCTVSGGSISSYYWSWIRQPPGKGLEWIGYIYYSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCASLPRGQLVNAYFDYWGQGTLVTVSS
I8              ----------------------------M----------S----------T-D-----------------E-------R-----------------------------------------
I4              ----------S---------------MGG---------S-V----------HT-D-------E----V------E-------R-----------------------F-------------
I3              ----------S---------------MGG-----L--S-V---------FHT-H-------E----V-----E-----R-R-----------------------F--N---------A-
I2              ----------V--S--------------MGGT----L-LS----------FHT-H-------EG--S------ED----R-R----------F--------------RN--R----S-TA
I1              ----------V--S--------------MGGT----L-LS----------FHT-H---S----G--S------ED----R-R----------F--------------RN--R----S-TA
I7              ---------------------S-----MT---------T----------T-D-----------------APE-------R-----------------------------------------
1AZCETI5        ---------------------S-----MTES--N-M-LT-TG--------FH--------A----------APG-H---R-R-L------T-F--R-R---------GS--P-I------
1AH92UCA        ----------S--------------MGG---------S-L----------HA-D-------E----V----------R-----------------------F--H----------
1A102RI6        ---------------------S-----MT----F----T----------T-E-H-------------APA-------R-------G--W---------IH---H-----A------
CH103           ----------V--S--------------MGGT----L-LS----------FHT-E---S----G--S------ED----R-R----------F--------------RN--R-S--S-TA
CH104           ----------V--S--------------MGGT----L-LS----------FHT-H-------EG--S------ED----R-R----------F--------------RN--R----S-TA
CH105           ----R---------S---------A-----MGG-----L-HS-V---------VFHT-H-------A---MV----TE-----RMR-----------------------F--N--P------A-
CH106           ----------V--S---------S-----MGGT----L-LS----------FHT-H---S----G--S------ED----R-R----------F--------------RN--R-I--S-TA

Light chain
                1                   23          35                  50     57      65                  80       89   95a  98
                |                   |           |                   |      |       |                   |        |    ||   |
                ---------FWR1---------___CDR1____------FWR2-----__CDR2_---------------FWR3-------------___CDR3___---FWR4--
                                      ○●☼                 ●☼☼☼                 ●●●○                 ●
UCA             SYELTQPPSVSVSPGQTASITCSGDKLGDKYACWYQQKPGQSPVLVIYQDSKRPSGIPERFSGSNSGNTATLTISGTQAMDEADYYCQAWDSFSTFVFGTGTKVTV
I2              ------------------T-------VAS-NV----V------EV-M-ENY-------D-----K--S------R----T--------V----------S--Q---
CH103           ------------------T-----A...STNV----V------EV--FENY-------D-----K--S------R----I--------V----------S--Q---
CH104           ------------------T-------VAS-NV----V------EV-M-ENY--A----D-----K--S------R----T--------V----------S--Q---
CH106           ------------------T-------VASVNV----V------DV-M-ENY-------D-----K--S------R----T---E----V----------S--Q---
```

Figure 7. **CH103 paratope, critical residues, and required immune precursors. a**, Overall structure of complex with variable domains of CH103 depicted in ribbon representation and gp120 shown as a molecular surface. The color scheme is the same as in Fig. 3a. **b**, CH103 paratope surface displayed on top of an underlying polypeptide ribbon. The surface is colored and labeled by contributing antibody components. **c**, CH103 paratope surface colored by maturation states of the underlying residues. Unmutated residues are colored magenta while affinity matured residues are colored green and light blue for heavy and light chains respectively. **d**, Sequence alignment of heavy and light chains of CH103 clonal lineage members. Framework and CDR residues are labeled, as are residues that interact with the gp120 (open circle, main chain interaction; open circle with rays, side chain interactions; filled circle, both main chain and side chain interactions). The unmutated paratope residues are highlighted in magenta and the maturation-gained paratope residues are highlighted in green for heavy chain and blue for light chain.

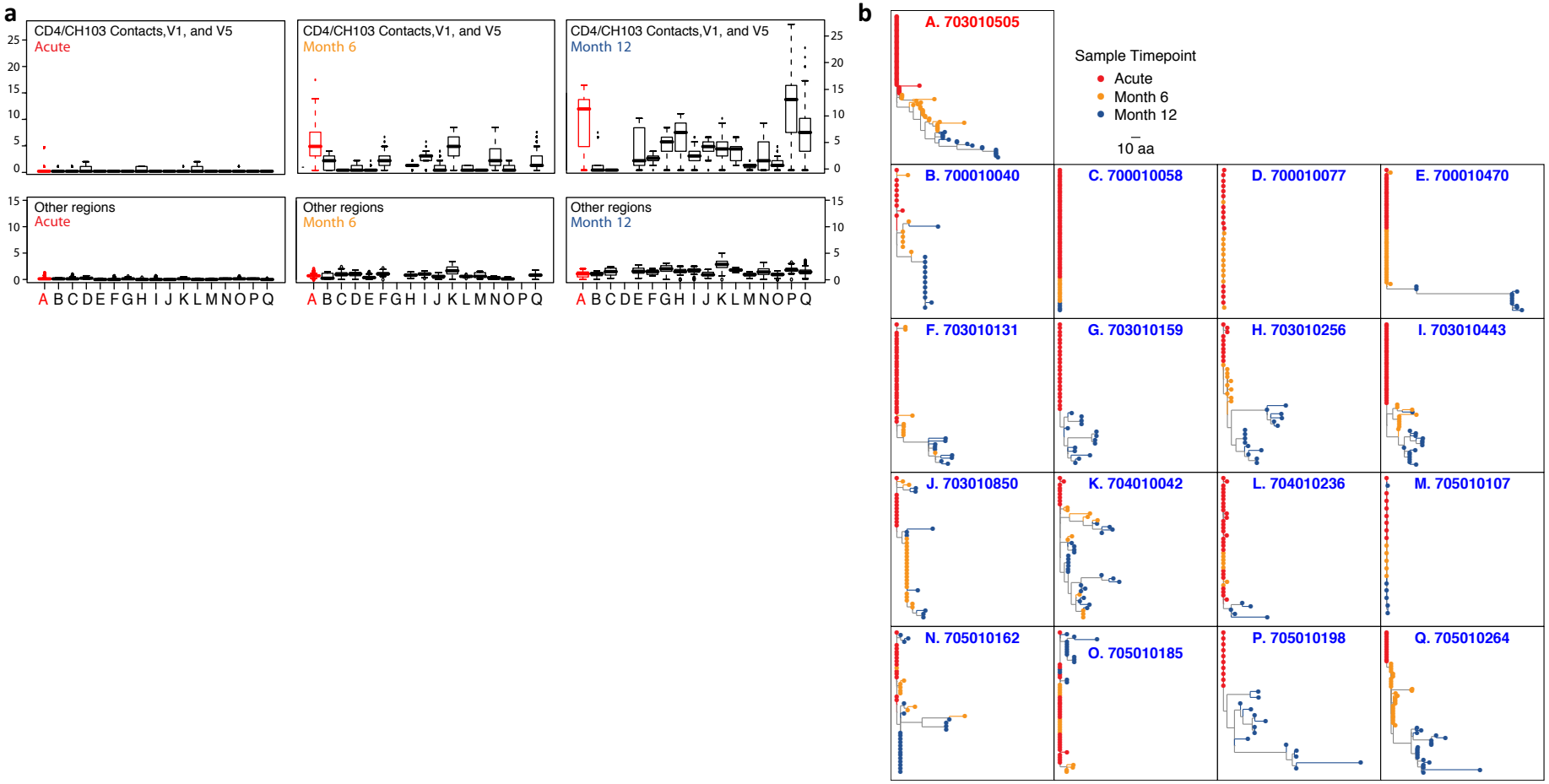


Supplementary Figure 8. Pixel map and phylogenetic tree of HIV-1 *env* gene evolution over time in CH505. The pixel tool (http://www.hiv.lanl.gov/content/sequence/pixel/pixel.html) was used to illustrate the amino acid changes in the V1 to V5 region of the envelope. we focused on this region as it is most critical for CD4bs antibody susceptibility, and includes all known CD4 binding contacts, which are indicated as black tic marks along the top of the figure. Blue tic marks indicated CH103 contact residues, and the horizontal blue line indicates that part of gp120 that was used for the CH103 crystal structure (although the contact surface is mostly there, still quite a bit is missing that is important for CD4 and VRC01, which is why we use CD4 contacts to help define bits that may be important for CH103 binding in those missing regions). Each row represents one sequence, and they are ordered according to the phylogeny. Red bits indicate amino acid changes relative to the TF virus, which was inferred as previously described [1,2]. Black bits indicate either an insertion or a deletion. The phylogenetic tree on the right was made with PhyML v2 [3] and the JTT substitution model [4] from the translated Env sequences. The tree was configured as a ladder and the T/F virus was reconstructed from the first time point sequences obtained at week 4 after transmission. Colors indicated the estimated number of weeks from infection. The tree was rendered with APE v3.0-6 [5] and both used R v2.15.1 [6]. The arrow indicates the week 30-53 selective bottleneck.

1. Keele BF et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc. Natl. Acad. Sci. USA, 105: 7552-7557 (2008).
2. Salazar-Gonzalez JF et al. Genetic identity, biological phenotype and evolutionary pathways of transmitted founder viruses in acute and early HIV-1 infection. J. Exp. Med. 206:1273-1289 (2009).
3. Guindon S, Gascuel O. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704 (2003).
4. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Applic Biosci 8: 275–282 (1992).
5. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290 (2004).
6. R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

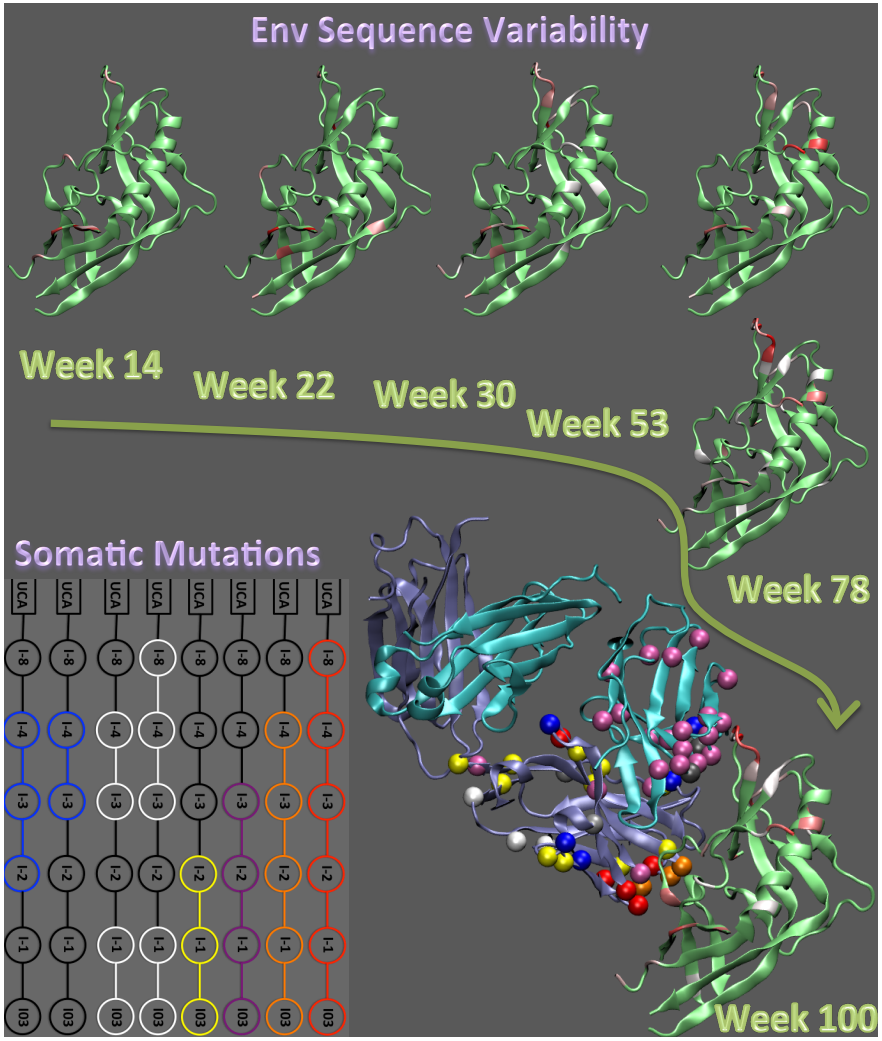


Supplementary Figure 9. Entropy map illustrating the per site diversity within each time point sampled in CH505. Full gp160 is shown, and CD4 and CH103 contact residues are highlighted. Shown is the Shannon entropy of each position in the alignment, where the observed frequency of all in a position characters is considered, and a gap is treated as a character (Korber J Virol. 1994;68(11):7467-81). This provides a map of regional within-time point diversity spanning Env, and illustrates where mutations are concentrated and the relative diversity of key regions over time.



Supplementary Figure 10. A comparison of the pace of viral sequence evolution in CH505 (indicated here by the 9-digit anonymous study-participant identifier 703010505) in regions relevant to the CH103 epitope with other subjects. The regions of interest include the CH103 contacts defined by the structure in this paper, as well as VRC01 contacts and CD4bs contacts[1], and the V1 and V5 loops immediately adjacent to these contacts. a, The distribution of sequence distances expressed as the percentage of amino acids that are different between two sequences, resulting from a pair-wise comparison of all sequences sampled in a given time point. Because these are all homogeneous (single-founder) infection cases, very few mutations appear in the CH103 relevant regions or other sites in the virus during acute infection (left hand panels). By 24 weeks after enrollment (week 30 from infection in (A) 703010505, labeled month 6 here as it is approximate), extensive mutations have begun to accrue, focused in CH103 relevant regions (top middle panel), but not in other regions of Env (bottom middle panel). Subject 703010505 has the highest ranked diversity among 15 subjects[2] (B-Q) sampled in this time frame (p=0.067), indicating a focused selective pressure began unusually early in this subject. By 1 year (month 12 indicates samples taken between 10-14 months from enrollment, due to variation in timing of patient visits), this region has begun to evolve in many individuals, possibly due to autologous NAb responses active later in infection. b, Phylogenetic trees based on concatenated CH103 relevant regions (HXB2 sites 124-127, 131, 132, 279-283, 364-371, 425-432, 455-465, 471-477) were created with PhyML3.0[3], using HIVw[4], a within-subject HIV protein substitution model, which was selected to be the optimum model for these sequences using ProtTest[5]. Indels were treated as an additional character state, rather than as missing information. In this view, the extensive evolution away from the T/F virus by month 6, shown in gold, is particularly striking. Distances between sequences sampled in 703010505 (A) at month 6 and the T/F ancestral state were significantly greater than the sequences in the next most variable individual (L) designated by the 9-digit identifier 704010042 (Wilcoxon rank sum, p = 0.0003: CH505, median = 0.064, range = 0.019 – 0.13, N = 25, and 704010042, median = 0.0271, range = 0.009 – 0.056, N = 26).

1. Zhou, T., *et al*. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* **329**, 811-817 (2010).
2. Liu MK, *et al*. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *J Clin Invest*. **123**, 380-93 (2013).
3. Guindon, S., *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**, 307-321 (2010).
4. Nickle, D.C., *et al.* HIV-specific probabilistic models of protein evolution. *PLoS ONE* **2**, e503 (2007).
5. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics **21**, 2104-2105 (2005).



Supplementary Figure 11. Co-evolution of virus and antibody - interplay between maturation of antibody CH103 and sequence variability epitope in gp120. The sequence variability (within sample) at each time point is mapped on a gp120 structure that tracks the viral evolution over time from 14 weeks through 100 weeks post-transmission. Entropy at each residue is color-coded as green to white to red to indicate no sequence variation to slight variation to high sequence variation. This extensive virus within-time point diversity coincides with maturing antibody lineages that ultimately develop breadth. Here we capture the somatic mutations along the CH103 clonal lineage beginning with unmutated common ancestor (UCA) to I-8 to I-4 to I-3 to I-2 to I-1 and to mature CH103. The color balls in heavy (violet) and light (cyan) chains of antibody indicate the appearance/disappearance of somatic mutations during the evolutionary path according to the following scheme. Red balls: mutations appeared in I-8 and remained all the way thru maturation to CH103, Orange balls: mutations appeared in I-4 and remained thru maturation to CH103, Blue balls: mutations that appeared early but are lost before maturing to CH103, and Gray balls: mutations that appeared very late in maturation. Structure of Fab CH103-gp120 from ZM176.66 complex determined in this work is used to map these mutations. The sequence entropy at week 100 is used to pair with CH103 to simply illustrate the relative spatial locations of somatic mutations and sequence variability in gp120. As discussed in the text, viral evolution with time tracks with neutralization breadth and this simple mapping supports that (i) T/F virus began to diversify very early in regions in or proximal to the epitope, and (ii) Somatic mutations that occur early in evolution and remain fixed in heavy chain tend to cluster near the gp120 contact region unlike those mutations that appear later.